ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	561168
ToLID	**puRebBill1**
Species	Rebecca
Class	NA
Order	Pavlovales

Genome Traits	Expected	Observed
Haploid size (bp)	168,125,043	72,036,265
Haploid Number	5 (source: ancestor)	34
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q58

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

. Observed Haploid size (bp) has >20% difference with Expected
. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed
. BUSCO single copy value is less than 90% for collapsed

Curator notes

. Interventions/Gb: 818
. Contamination notes: ""
. Other observations: "The assembly of Rebecca billardiae RCC1528 (puRebBill1.1) is based on 40X PacBio data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio assembly generation with Hifiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 361 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 34.8 Mb (with the largest being 5.91 Mb). Additionally, 285 regions totaling 20.4 Mb (with the largest being 0.77 Mb) were identified as haplotypic duplications and removed. The mitochondrial and chloroplastic genomes were assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 10 haplotypic regions and 71 contaminant sequences were removed, totaling 0.99 Mb and 0.77 Mb (with the largest being 0.33 Mb and 0.04 Mb). Chromosome 19 has a different coverage pattern, suggesting a large heterozygous deletion of several hundred

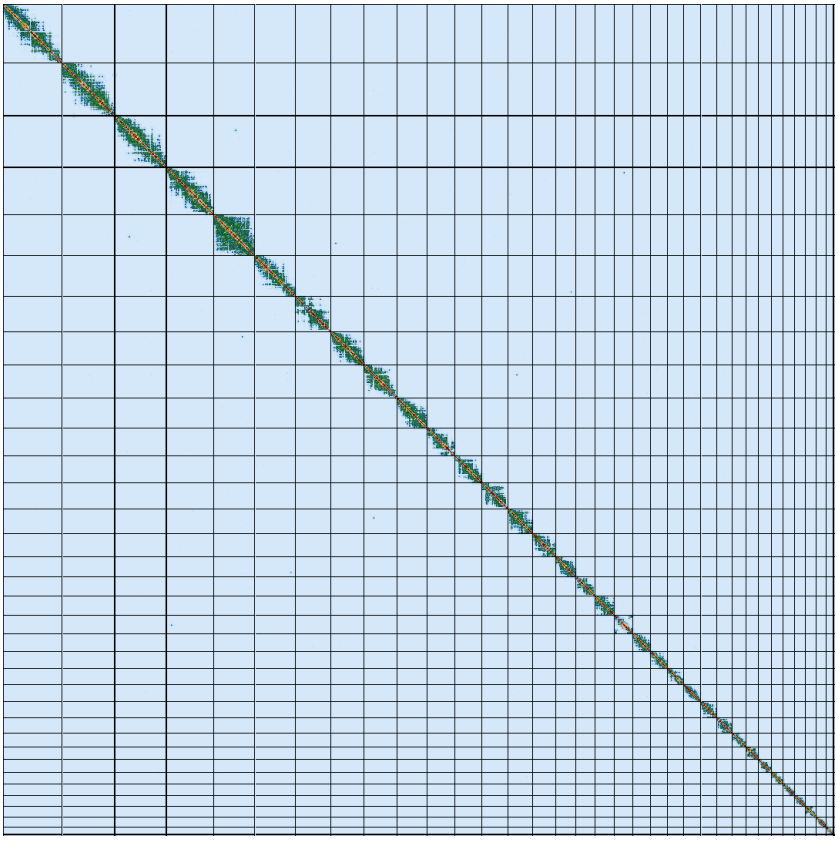
kilobases. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	73,905,732	72,036,265
GC %	69.77	70.13
Gaps/Gbp	784.78	916.21
Total gap bp	5,800	8,900
Scaffolds	119	41
Scaffold N50	2,846,454	2,611,987
Scaffold L50	10	10
Scaffold L90	26	27
Contigs	177	107
Contig N50	1,157,098	1,150,000
Contig L50	21	20
Contig L90	65	64
QV	34.0207	58.4558
Kmer compl.	46.5124	68.5797
BUSCO sing.	72.5%	72.9%
BUSCO dupl.	1.6%	0.8%
BUSCO frag.	8.2%	8.2%
BUSCO miss.	17.7%	18.1%

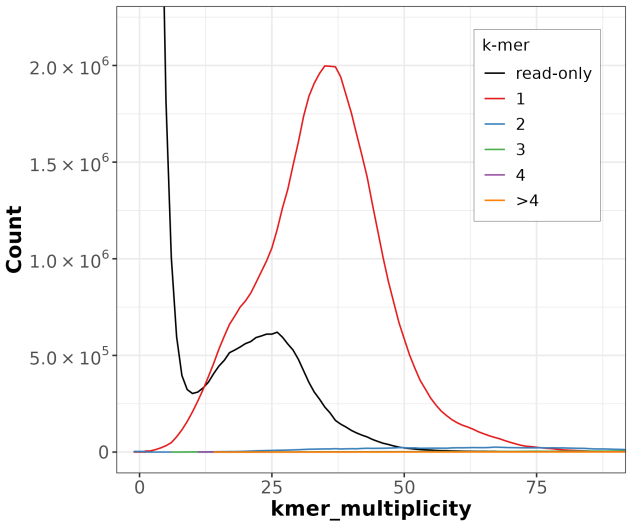
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: eukaryota_odb10 (genomes:70, BUSCOs:255)

HiC contact map of curated assembly

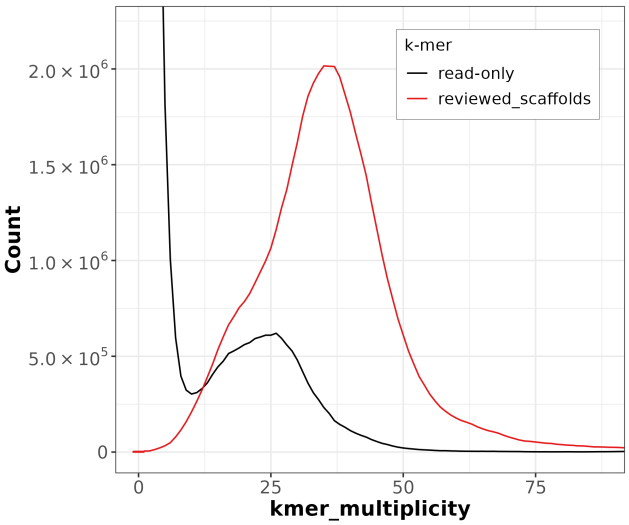


collapsed [LINK]

K-mer spectra of curated assembly

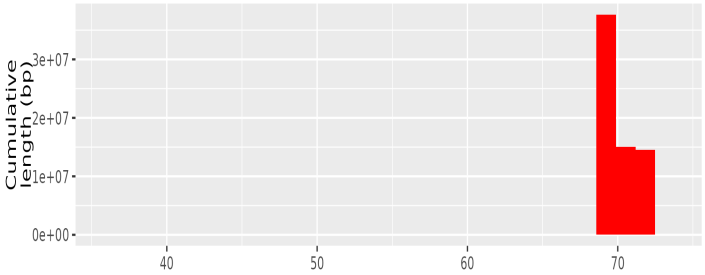


Distribution of k-mer counts per copy numbers found in asm

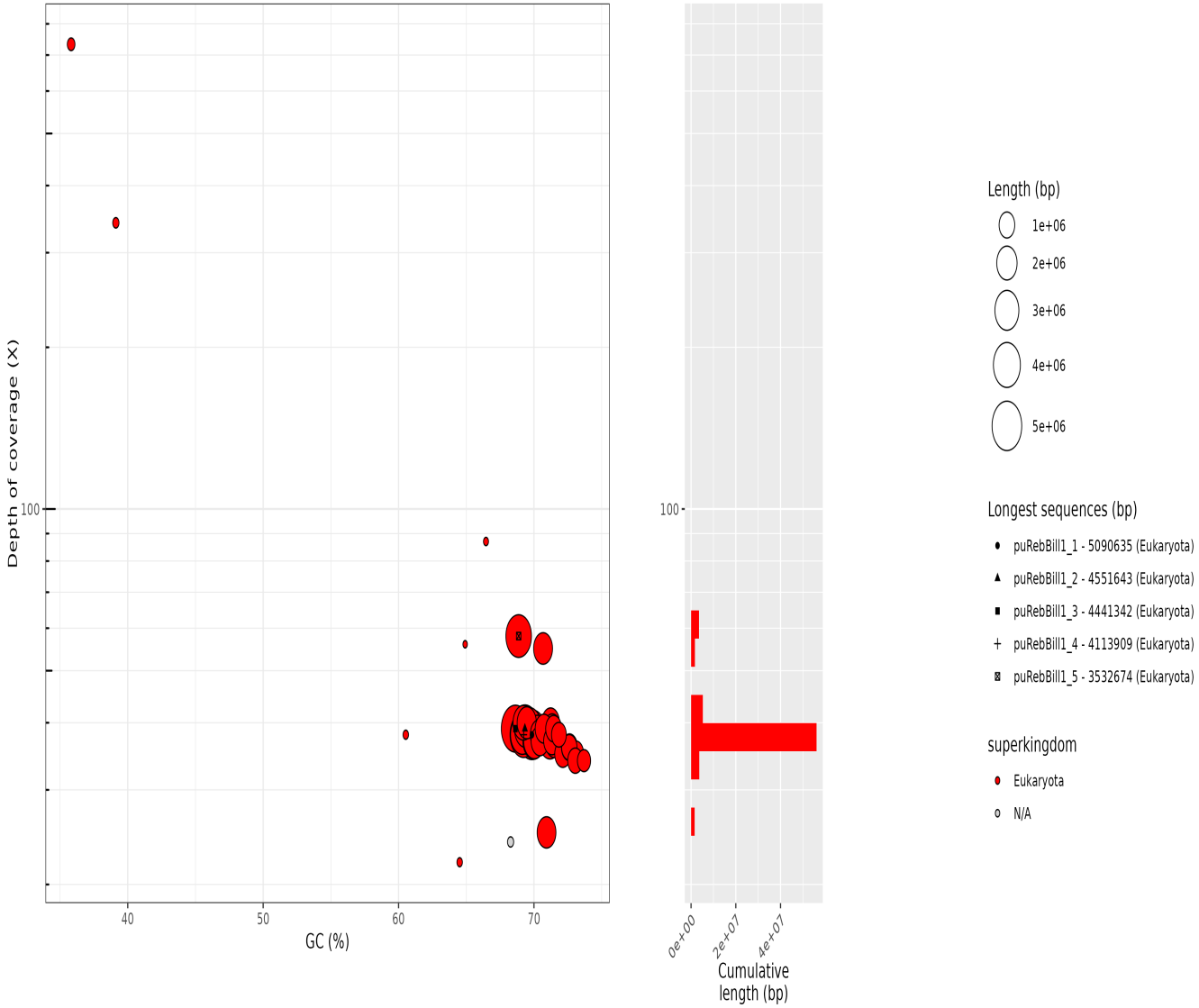


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



collapsed. Bubble plot circles are scaled by sequence length, positioned by coverage and GC proportion, and coloured by taxonomy. Histograms show total assembly length distribution on each axis.

Data profile

Data	PACBIO Hifi	Arima
Coverage	40	158

Assembly pipeline

- **Hifiasm**
  |_ *ver:* 0.19.5-r593
  |_ *key param:* NA
- **purge_dups**
  |_ *ver:* 1.2.5
  |_ *key param:* NA
- **YaHS**
  |_ *ver:* 1.2
  |_ *key param:* NA

Curation pipeline

- **PretextMap**
  |_ *ver:* 0.1.9
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 0.2.5
  |_ *key param:* NA

Submitter: Jean-Marc Aury
Affiliation: Genoscope

Date and time: 2025-04-05 14:47:09 CEST